# Overview

RunPod offers Serverless GPU computing for AI inference, training, and general compute, allowing users to pay by the second for their compute usage. This flexible platform is designed to scale dynamically, meeting the computational needs of AI workloads from the smallest to the largest scales.

You can use the following methods:

- Quick Deploy: Quick deploys are pre-built custom endpoints of the most popular AI models.
- Handler Functions: Bring your own functions and run in the cloud.

If you'd like to use pre-built endpoints of some of the most popular AI models, see AI APIs.

## Why RunPod Serverless?

You should choose RunPod Serverless instances for the following reasons:

- **AI Inference:** Handle millions of inference requests daily and can be scaled to handle billions, making it an ideal solution for machine learning inference tasks. This allows users to scale their machine learning inference while keeping costs low.

- **AI Training:** Machine learning training tasks that can take up to 12 hours. GPUs can be spun up per request and scaled down once the task is done, providing a flexible solution for AI training needs.

- **Autoscale:** Dynamically scale workers from 0 to 100 on the Secure Cloud platform, which is highly available and distributed globally. This provides users with the computational resources exactly when needed.

- **Container Support:** Bring any Docker container to RunPod. Both public and private image repositories are supported, allowing users to configure their environment exactly how they want.

- **3s Cold-Start:** To help reduce cold-start times, RunPod proactively pre-warms workers. The total start time will vary based on the runtime, but for stable diffusion, the total start time is 3 seconds cold-start plus 5 seconds runtime.

- **Metrics and Debugging:** Transparency is vital in debugging. RunPod provides access to GPU, CPU, Memory, and other metrics to help users understand their computational workloads. Full debugging capabilities for workers through logs and SSH are also available, with a web terminal for even easier access.

- **Webhooks:** Users can leverage webhooks to get data output as soon as a request is done. Data is pushed directly to the user's Webhook API, providing instant access to results.

RunPod Serverless GPUs are not just for AI Inference and Training. They're also great for a variety of other use cases. Feel free to use them for tasks like rendering, molecular dynamics, or any other computational task that suits your fancy.

## How to interact with RunPod Serverless?

RunPod generates an Endpoint Id that that allows you to interact with your Serverless Pod. Pass in your Endpoint Id to the Endpoint URL and provide an operation.

This Endpoint URL will look like this:

```
https://api.runpod.ai/v2/{endpoint_id}/{operation}
```

- `api.runpod.ai`: The base URL to access RunPod.
- `v2`: The API version.
- `endpoint_id`: The ID of the Serverless Endpoint.
- `operation`: The operation to perform on the Serverless Endpoint.
  - Valid options: `run` | `runsync` | `status` | `cancel` | `health` | `purge`

Edit this page